ECE 517: Machine Learning

Assignment 2.1: Minimum Mean Square Criterion

David Kirby – 101652098 – davidkirby@unm.edu

Fall 2021

SCHOOL OF ENGINEERING

A variation of the MMSE criterion minimizes the norm of the weight vector **w**. This is a way to control the complexity of the structure. The corresponding function is

$$\mathcal{L}(\mathbf{x}, \mathbf{w}) = \mathbb{E}\left[e^2\right] + \lambda \parallel \mathbf{w} \parallel^2$$

1. Make the derivation of the closed solution for **w**.
2. Work out an iterative solution using the same technique as used in the Least Mean Squares algorithm.
3. Comment and compare both solutions in a short conclusion section.

The derivations must be complete and the solution should be briefly but completely explained. See the rubric for this and any other homework.

1. Make the derivation of the closed solution for **w**.

$$\begin{aligned}
\mathcal{L}(\mathbf{w}) &= \mathbb{E}\left[e^2\right] + \lambda \parallel \mathbf{w} \parallel^2 \\
&= \frac{1}{N}\sum_{i=1}^{D}\left(y_n - \mathbf{w}^\top \mathbf{x}_n\right)^2 + \lambda \mathbf{w}^\top \mathbf{w} \qquad (1) \\
&= \sum_{i=1}^{D} y_n^2 + \mathbf{w}^\top \mathbf{X}\mathbf{X}^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{X}\mathbf{y} + \lambda \mathbf{w}^\top \mathbf{w} \\
&= \sum_{i=1}^{D} y_n^2 + \mathbf{w}^\top \mathbf{R}\mathbf{w} - 2\mathbf{w}^\top \mathbf{p} + \lambda \mathbf{w}^\top \mathbf{w} \\
\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}) &= 2\mathbf{R}\mathbf{w} - 2\mathbf{p} + 2\lambda \mathbf{w} = 0 \qquad (2) \\
&= \mathbf{R}\mathbf{w} - \mathbf{p} + \lambda \mathbf{w} \\
&= \mathbf{R}\mathbf{w} - \mathbf{p} + \lambda \mathbf{I}\mathbf{w} \\
\mathbf{w} &= [\mathbf{R} + \lambda \mathbf{I}]^{-1}\mathbf{p} \qquad (3)
\end{aligned}$$

Equation (1) incorporates a structure and a training criterion. For our structure, we can use a family of linear functions: $\hat{y_n} = \mathbf{w}^\top \mathbf{x}_n + b$. A criterion then needs to be chosen to optimize parameters $\mathbf{w}$. The simplest criterion in supervised learning is the minimization of the mean square error: $e^2 = (y_n - \hat{y_n})^2$. We can then use the law of large numbers to approximate. The only thing left to do is to expand the polynomial and replace variables with $\mathbf{R}$ and $\mathbf{p}$, which are estimates of the data autocorrelation matrix and the cross correlation vector between data and labels, respectively. Equation (2) is the optimization performed by computing the gradient with respect to $\mathbf{w}$, which is then nullified. Finally, solving for $\mathbf{w}$ results in equation (3), a set of Widrow–Hoff equations.

2. Work out an iterative solution using the same technique as used in the Least Mean Squares algorithm.

$$\begin{aligned}\mathbf{w}^{k+1} &= \mathbf{w}^k - \mu\Big[\mathbf{R}\mathbf{w}^k - \mathbf{p} + \lambda\mathbf{I}\mathbf{w}^k\Big] && (4)\\ &= \mathbf{w}^k - \mu\left[\mathbf{x}_n\mathbf{x}_n^\top\mathbf{w}^k - \mathbf{x}_n y_n\right]\\ &= \mathbf{w}^k - \mu e_n\mathbf{x}_n && (5)\end{aligned}$$

With a steepest descent procedure, $\mathbf{w}$ is iteratively optimized by changing it in the direction opposite to its gradient. $\mathbf{R}$ is again an estimate of the data autocorrelation matrix with $\mathbf{R} \approx \mathbf{x}_n\mathbf{x}_n^\top$ and $\mathbf{p}$ is the cross correlation vector between data and labels with $\mathbf{p} \approx \mathbf{x}_n y_n$. To simplify, μ is a small parameter and e_n is the estimation error.

3. Comment and compare both solutions in a short conclusion section.

Equation (5) is an optimization that updates the parameters of equation (3) toward a maximum descent of the gradient. This equation, the Least Mean Squares Algorithm, is the least computationally burdensome, but, as the professor says, there is no free lunch. Least Mean Squares can have issues with samples that come one at a time, or with computationally complex $\mathbf{R}$ and $\mathbf{p}$.